DTIC/TR-89/20

AD-A214 147

AD-A214 147

DAITC

TERMINOLOGY STRATEGIES FOR INTERNATIONAL INFORMATION EXCHANGE

DAITC/TR-89/9

Gladys A. Cotter
Director, Defense Applied Information Technology Center

Walter R. Blados
Deputy for Scientific and Technical Information, Pentagon, USAF

August 1989

Defense Applied Information Technology Center

1800 North Beauregard Street
Alexandria, Virginia 22311
(703) 998-4787
Fax No. (703) 931-3968

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) DAITC/TR-89/9	5. MONITORING ORGANIZATION REPORT NUMBER(S) DTIC/TR-89/20
6a. NAME OF PERFORMING ORGANIZATION Defense Applied Information Technology Center 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Defense Technical Information Center
6c. ADDRESS (City, State, and ZIP Code) 1800 N. Beauregard Street Alexandria, VA 22311	7b. ADDRESS (City, State, and ZIP Code) Cameron Station Alexandria, VA 22304-6145
8a. NAME OF FUNDING/SPONSORING ORGANIZATION USAF Deputy for Scientific & Technical Info. 8b. OFFICE SYMBOL (If applicable) SAF/AQT	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) The Pentagon Washington, DC 20330	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Terminology Strategies for International Information Exchange (Unclassified)

12. PERSONAL AUTHOR(S)
Gladys A. Cotter, Walter R. Blados

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
	FROM _____ TO _____	8908	13

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
5	2		Thesauri, Core thesauri, Bilingual thesauri, Vocabulary, Micro-vocabularies, Scientific and technical information, (continued on reverse)
12	5		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Development of a common thesaurus for use by national and international Defense Scientific and Technical Information (STI) organizations to facilitate the exchange of information is described. The concept of a "core" thesaurus comprising STI terminology which is acceptable to all participants and which can be extended with specific micro vocabularies at the local level to meet specialized needs is explored. The thesaurus will be bilingual with both English and French terminology.

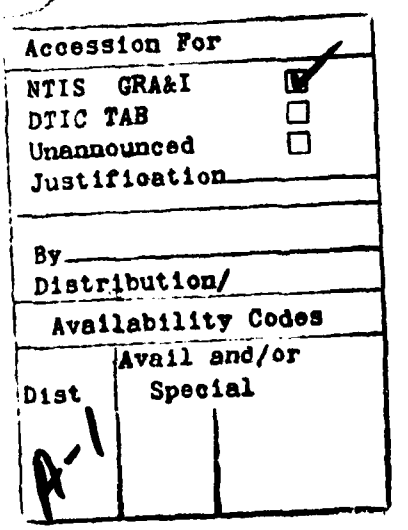
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Gladys A. Cotter	22b. TELEPHONE (Include Area Code) (703) 845-9400 / 22c. OFFICE SYMBOL DTIC-DA

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

Block 18. *SUBJECT TERMS*

STI, English terminology, French terminology, North Atlantic Treaty Organization, NATO



TERMINOLOGY STRATEGIES FOR INTERNATIONAL INFORMATION EXCHANGE

Gladys A. Cotter
Director, Defense Applied Information Technology Center

Walter R. Blados
Deputy for Scientific and Technical Information, Pentagon, USAF

August 1989

Defense Applied Information Technology Center
Alexandria, VA 22312

United States Air Force, Deputy for Scientific and Technical Information
The Pentagon, Washington, D.C. 20330

TERMINOLOGY STRATEGIES FOR INTERNATIONAL INFORMATION EXCHANGE

Gladys A. Cotter
Walter R. Blados

Key Words: Thesauri, core thesauri, bilingual thesauri, vocabularies, micro-vocabularies, scientific and technical information, STI, English terminology, French terminology, North Atlantic Treaty Organization, NATO.

Abstract: Development of a common thesaurus for use by national and international Defense Scientific and Technical Information (STI) organizations to facilitate the exchange of information is described. The concept of a "core" thesaurus comprising STI terminology which is acceptable to all participants and which can be extended with specific micro vocabularies at the local level to meet specialized needs is explored. The thesaurus will be bilingual with both English and French terminology.

I. INTRODUCTION

The scientific and the technical information (STI) community is a diverse body with a global pulse. The STI membership is segmented by subject, organization, language, and national security boundaries. Access to pertinent information regardless of source is vital to this community. Results of a study performed by one organization can accelerate the progress of another organization or allow them to avoid an unnecessary expenditure. Many STI organizations share the common goal to develop techniques to increase opportunities for exchange of STI among different organizations on a global basis.

This paper documents an effort to develop a common thesaurus for use by national and international STI organizations to facilitate the exchange of information.

The effort involves the development of a "core" thesaurus comprising STI terminology which is acceptable to all participants. Conceptually, the core thesaurus is extendable at the local level through specific micro vocabularies developed to meet specialized needs falling outside of the common core. All micro vacabulary hierarchies contain a top term from the core thesaurus, thus providing an integrated vocabulary structure at both the local and international levels. The thesaurus, when complete, will be available initially in English and French.

The immediate application of the thesaurus is as a tool for document indexing to facilitate the exchange of bibliographic information among the participating organizations. The thesaurus provides, however, the foundation for developing terminology related to knowledge bases which will be necessary for full text retrieval systems, natural language and expert systems retrieval aids, and computer indexing strategies which are currently being planned. The thesaurus will be distributed on magnetic tape and CD-ROM.

II. APPROACH

A committee composed of representatives from Belgium, Canada, the Federal Republic of Germany, France, The Netherlands, the United Kingdom and the United States was established under sponsorship of the North Atlantic Treaty Organization (NATO). The task of the committee was to develop a terminology strategy for facilitating retrieval of information from databases by members of the international community in order to avoid duplication of effort. The terminology structure/system developed by the committee will be incorporated in the design of a centralized database at NATO. In addition, the system will facilitate the exchange of information among national and international databases at the discretion of the nations. The existing national databases and the NATO database being created consist of, primarily, citation data referencing full text documents. In most cases, the database owners plan to move to online full text storage as soon as economically and technologically feasible.

The committee was given requirements which had to be incorporated in the selected strategy. The requirements were that the solution should:

a. Accommodate both English and French and be fully bilingual.
b. Be easy to use by end users and not require information specialists to retrieve accurate results.
c. Allow for full cross relationships in order to identify related topics, not merely indentical topics.
d. Be of comprehensive scope to meet the needs of international interests.
e. Allow for growth and adaptability so that it could evolve and meet changing user requirements.
f. Allow for the fundamental capabilities to be implemented within a 2-year time frame.

Given these requirements, the group set out to identify alternative approaches. The alternatives selected for review were:

a. Utilization of a database management system (DBMS) with full text retrieval.
b. Utilization of a categorization system.
c. Utilization of a thesaurus/controlled vocabulary.

The DBMS full text retrieval option was attractive because of the low cost of input (indexers were not needed) and the ease of use. Many of the committee members were using this approach with some of their national databases and had experienced satisfactory results. The difficulty with this type of system is that important information can be missed because the searcher and the author may use different terminology to describe the same concept. Therefore, the system must be built with a synonym directory (including specialized scientific and technical terminology), or the searcher must spend a considerable amount of time developing the search strategy (offsetting the ease of use and low cost of input). These problems may be relatively simple to overcome when only a single language is involved. In this case, dealing with two languages and with the fact that many of the users have a native language other than English or French adds another dimension of complexity. The group decided that the full text retrieval option as a stand-alone solution was not optimal for the application under consideration.

The next possibility reviewed was utilization of a categorization scheme such as the COSATI (Committee on Scientific and Technical Information) Fields and Groups, the Subject Categorization Guide for Defense Science and Technology Subject Categorization Guide for Defense Science and Technology or the Universal Decimal Code (UDC). The categorization schemes did not allow for sufficient cross relationships to meet the requirements. It also appeared that it would be difficult to expand and "grow" most of the categorization schemes on the scale that would be required for this effort. This solution was rejected by the committee.

A thesaurus was the next option considered. The committee noted that a fully developed thesaurus targeted for this purpose would provide for full cross relationships, a comprehensive scope, growth potential and ease of use. The thesaurus would also eliminate some of the full text searching problems such as different spellings and use of synonyms and homonyms. In addition, it could be used without an abstract and makes up for the lack of precision in title terminology. On the negative side, a thesaurus can be a costly and labor intensive system to develop, maintain and use. We decided that the thesaurus option should be further explored to determine if an existing thesaurus would meet the requirements, if one could be used as a baseline and adapted to meet our needs, or if a new thesaurus would have to be created. Each member of the group was tasked to review thesauri utilized in their country and make recommendations regarding suitable thesauri. One of the major constraints on selecting candidates was that military terminology should be either included already or be easily incorporated into any thesaurus selected. Nine systems were selected for evaluation against the following criteria:

a. Scope of coverage.
b. Adaptability of the structure, including hierarchiacal, alphabetical, and key word out of context (KWOC) structures.
c. Polyhierarchical and polydimensional structure of the concepts.
d. Availability of cross relations, including an equivalence relation (such as USE), a hierarchical relation (such as narrower term (NT) and broader term (BT) and an associative relation (such as related term (RT)).
e. User friendliness, including ease of use and incorporation of definitions, scope notes, and cross-references.
f. Military terminology and weapon systems nomenclature.
g. Ease of maintenance and expansion.
h. Bilingual availability (English and French).

The committee made theoretical evaluations of the nine thesauri and performed practical tests involving indexing representative documents using each of the thesauri.

The result was a two-fold determination: that (1) no existing thesaurus fully met the requirements; (2) the Defense Technical Information Center (DTIC) Retrieval and Indexing Terminology (DRIT) could be adapted to meet the requirements. Adaptation of the DRIT to meet the requirements involved extending the exiting military terminology, implementing associated relations (related terms), developing a French version, implementing bilingual (French/English) concept relationships, and deleting terminology of a national as opposed to international orientation.

The group also decided to use the field and group categorization scheme associated with DRIT, the Subject Categorization Guide for Defense Science and Technology.

The group felt that a baseline thesaurus consisting of the DRIT adapted to include bilingual terminology and additional military terminology could be made available within a 2-year timeframe. They also felt that the obvious national terminology could be deleted with the 2-year timeframe. Introduction of related terms into the DRIT baseline was recognized as a long-term task requiring 3 to 6 years. Maintenance of the thesaurus, including adding new tems as documents are indexed and as the shortcomings of the baseline become apparent, would be a continuing task.

The group selected the approach of developing a thesaurus adapted from the DRIT. They recommended that the thesaurus be utilized in conjunction with a DBMS that provided full text retrieval capabilities. It was felt that by combining these capabilities, the system users would have a powerful and flexible retrieval capability.

III. BASELINE THESAURUS IMPLEMENTATION

The first task was to expand the DRIT military terminology. Each of the comittee members was assigned a subject area. The members were asked to review the terminology and make recommendations for additions and deletions. For new terminology, placement in the heirarchy also had to be addressed. The translation task involved translating concepts not merely words.

This task was initiated by obtaining a machine-readable tape of the DRIT baseline and processing it through the SYSTRAN Translation Systems Inc. This action resulted in a "raw" translation requiring human review to ensure that similar concepts were represented accurately in both the English and French versions. The representatives from Canada and France provided resources for performing the review of the machine translation.

The software packages used to create the online baseline thesaurus were Inquire Text and AVOCON from Infodata Systems Inc. A machine-readable tape of the DRIT was loaded as a first step in creating the online thesaurus. New terminology currently being developed will be added as it becomes available. The online thesaurus design calls for the French and English versions of the thesaurus to be maintained in separate files with linkages between related terminology and concepts.

The baseline thesaurus will be available for use online in early 1990.

IV. NETWORK STRATEGIES

The committee task was to develop a plan for a terminology system that could be adopted for use by member nations as well as implemented in the central location. The effort to date has been focused primarily on the centralized implementation. A network strategy has been considered, and several plans are being formulated, the rudiments of which are described below. In a network strategy, the thesaurus would initially be distributed on magnetic tape; however, distribution on CD-ROM is planned as a near-term enhancement.

When the committee considered potential large scale use of the thesaurus by multiple, diverse user groups, it became obvious that trying to develop a single thesaurus to serve such a large disparate audience had many problems. A thesaurus acceptable to everyone would be so broad as to dilute its purpose. Developing a thesaurus specific enough to cover all needs would be an impossible task in any type of reasonable timeframe.

One option considered was to provide the thesaurus and updates to requesting organizations and allow them to submit requests for new terminology to a central authority. It their suggestions were accepted, those suggestions would be integrated into the structure and distributed with the next update.

If their suggestions were rejected, they would have the option of not using the teminology or implementing it anyway and deviating from the standard.

The committee considered this solution to be unsatisfactory for several reasons. First, they were concerned that most terminology would have to be rejected in order to maintain a manageable central thesaurus. Second, no mechanism would exist to share information regarding their addition of terminology to the local versions of the thesaurus. For example, if one organization added terminology on "animal husbandry" that was not included in the central version, a second organization which had a need for terminology in that same area would have no opportunity to implement the structure already created.

The idea of a centrally controlled "core" thesaurus which could be extended with specific micro vocabularies at the local level to meet specialized needs falling outside of the common core was thought to be more advantageous. An organization wishing to add new terminology to meet local needs would be asked to select a term from the "core" thesaurus to form the top term of the local structure. In this manner, linkages would be built between the local and the core thesaurus. Organizations building micros for local use would inform the central authority of the subject area. The central authority would make this information available to other users of the "core" thesaurus.

Another way the core thesaurus could be used to facilitate the exchange of information is through agreements reached between interested organizations to use a certain percentage of core terms in addition to their own local terminology, when indexing documents. With these commonalities, participants could integrate other's information into their database collections without re-indexing. They could also select information for exchange based on subject profiles using core terminology.

V. FUTURE D'RECTIONS

As the committee developed the strategy for implementing and utilizing the thesaurus, one discussion centered on whether expet systems would make a thesaurus obsolete in the near future.

The concern was that resources would be invested in development of a system that would be overtaken by technological advances before a reasonable return on investment could be achieved. If this were the case, it would be wiser to rely on the DBMS with full text retrieval capabilities for the present and wait for improvements in natural language searching. Users would prefer to use an efficient natural language retrieval system, if available, rather than a retrieval system based on a controlled vocabulary.

We decided to explore the state of artificial intelligence and expert system technologies required to provide a natural language solution for our requirement. The operational natural language retrieval and indexing systems we identified were all limited to a specific, manageable area of interest. Unfortunately, neither current technology nor predicted operational technology available in the next 7 years can address our broad application across multiple areas of interest in defense, science and technology.

We also found that natural language expert systems require that knowledge bases be formulated to support the function. Knowledge bases consist of terminology, usage contexts, related terminology and concepts, and syntax formats. It became apparent that the thesaurus being developed could be considered the beginning of a knowledge base; we were developing the common terminology, relationships and definitions that a knowledge base requires. As local micro vocabularies are developed, these will enrich the knowledge base foundation.

Our original task was to develop a useful system which would be implemented within 2 years and would allow for growth and adaptability as requirements change. The thesaurus solution being implemented meets the 2-year requirement. It can also serve as a foundation for future retrieval systems utilizing natural language and expert systems. Further, terminology in the thesaurus can be built into computer indexing strategies.

Enhancing retrieval and indexing capabilities will become more important as full text of documents, rather than merely citations to documents, are available online. Having full text available will magnify the problem of culling useful information from the massive store of available data.

It is our hope that the thesaurus we are developing can play an important transition role in the future as well as satisfy immediate, present day needs to facilitate information retrieval within the national and international context.